

AUTOMATIC COMPLETION OF FRAGMENTS OF TEXT

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates generally to information retrieval systems and, more particularly, to systems and methods for automatically completing fragments of text (e.g., sentences or paragraphs).

Description of Related Art

[0002] Oftentimes, people have trouble completing sentences and/or paragraphs. They know what they want to say but they cannot find the appropriate words to say it. These people may find it beneficial to be offered possible completions for sentences and/or paragraphs.

[0003] Accordingly, there exists a need for mechanisms that provide possible completions for fragments of text, such as partial sentences and/or paragraphs.

SUMMARY OF THE INVENTION

[0004] Systems and methods, consistent with the principles of the invention, automatically complete fragments of text, such as sentences or paragraphs.

[0005] According to one aspect consistent with the principles of the invention, a method for completing fragments of text is provided. The method may include obtaining a text fragment and performing a search, based at least in part on the text fragment, to identify one or more documents. The method may also include identifying sentences within the one or more documents that are associated with the text fragment, determining sentence endings associated with the identified sentences, and presenting the sentence endings as potential completions for the text fragment.

[0006] According to another aspect, a computer device includes a memory configured to store code and a processor configured to execute the code in the memory. The code in the memory may include document preparation code and assistant code. The document preparation code is configured to permit a user to prepare or edit a document. The assistant code is configured to detect a fragment of text within the document, obtain potential sentence completions for the fragment of text, and present the potential sentence completions to the user.

[0007] According to a further aspect, a computer device includes a memory configured to store instructions and a processor configured to execute the instructions in the memory. The processor may obtain a fragment of text and search for local documents that include at least a portion of the fragment of text. The processor may identify sentences within the local documents that are associated with the fragment of text, determine sentence completions associated with the located sentences, and provide the sentence completions as potential completions for the fragment of text.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

[0009] Fig. 1 is a diagram of an exemplary network in which systems and methods consistent with the principles of the invention may be implemented;

[0010] Fig. 2 is an exemplary diagram of a client and/or server of Fig. 1 in an implementation consistent with the principles of the invention;

[0011] Figs. 3A and 3B are flowcharts of exemplary processing for automatically completing a fragment of text according to an implementation consistent with the principles of the invention; and

[0012] Fig. 4 is a diagram of an exemplary ranked list according to an implementation consistent with the principles of the invention.

DETAILED DESCRIPTION

[0013] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention.

[0014] Systems and methods consistent with the principles of the invention may automatically complete a fragment of text, such as a sentence or paragraph. The systems and methods may identify possible endings from documents, such as web documents, and provide these endings as possible completions for the fragment of text.

EXEMPLARY NETWORK CONFIGURATION

[0015] Fig. 1 is an exemplary diagram of a network 100 in which systems and methods consistent with the principles of the invention may be implemented. Network 100 may include multiple clients 110 connected to multiple servers 120-140 via a network 150. Network 150 may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, a memory device, another type of network, or a combination of networks. Two clients 110 and three servers 120-140 have been illustrated as connected to network 150 for simplicity. In practice, there may be

more or fewer clients and servers. Also, in some instances, a client may perform the functions of a server and a server may perform the functions of a client.

[0016] Clients 110 may include client entities. An entity may be defined as a device, such as a wireless telephone, a personal computer, a personal digital assistant (PDA), a lap top, or another type of computation or communication device, a thread or process running on one of these devices, and/or an object executable by one of these device. Servers 120-140 may include server entities that gather, process, search, and/or maintain documents in a manner consistent with the principles of the invention. Clients 110 and servers 120-140 may connect to network 150 via wired, wireless, and/or optical connections.

[0017] In an implementation consistent with the principles of the invention, server 120 may optionally include a search engine 125 usable by clients 110. Server 120 may crawl a corpus of documents (e.g., web pages) and store information associated with these documents in a repository of crawled documents. Servers 130 and 140 may store or maintain documents that may be crawled by server 120. While servers 120-140 are shown as separate entities, it may be possible for one or more of servers 120-140 to perform one or more of the functions of another one or more of servers 120-140. For example, it may be possible that two or more of servers 120-140 are implemented as a single server. It may also be possible for a single one of servers 120-140 to be implemented as two or more separate (and possibly distributed) devices.

EXEMPLARY CLIENT/SERVER ARCHITECTURE

[0018] Fig. 2 is an exemplary diagram of a client or server entity (hereinafter called "client/server entity"), which may correspond to one or more of clients 110 and servers 120-140, according to an implementation consistent with the principles of the invention. The client/server entity may include a bus 210, a processor 220, a main memory 230, a read only memory (ROM)

240, a storage device 250, one or more input devices 260, one or more output devices 270, and a communication interface 280. Bus 210 may include one or more conductors that permit communication among the components of the client/server entity.

[0019] Processor 220 may include any type of conventional processor or microprocessor that interprets and executes instructions. Main memory 230 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by processor 220. ROM 240 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 220. Storage device 250 may include a magnetic and/or optical recording medium and its corresponding drive.

[0020] Input device(s) 260 may include one or more conventional mechanisms that permit an operator to input information to the client/server entity, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. Output device(s) 270 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, a speaker, etc. Communication interface 280 may include any transceiver-like mechanism that enables the client/server entity to communicate with other devices and/or systems. For example, communication interface 280 may include mechanisms for communicating with another device or system via a network, such as network 150.

[0021] As will be described in detail below, the client/server entity, consistent with the principles of the invention, perform certain searching-related operations. The client/server entity may perform these operations in response to processor 220 executing software instructions contained in a computer-readable medium, such as memory 230. A computer-readable medium may be defined as one or more physical or logical memory devices and/or carrier waves.

[0022] The software instructions may be read into memory 230 from another computer-readable medium, such as data storage device 250, or from another device via communication interface 280. The software instructions contained in memory 230 causes processor 220 to perform processes that will be described later. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the principles of the invention. Thus, implementations consistent with the principles of the invention are not limited to any specific combination of hardware circuitry and software.

EXEMPLARY PROCESSING

[0023] Figs. 3A and 3B are flowcharts of exemplary processing for automatically completing fragments of text, such as sentences and paragraphs, according to an implementation consistent with the principles of the invention. Processing may begin with server 120 receiving a search query from a user (act 310) (Fig. 3A). For example, a user may use conventional web browser software on client 110 to access search engine 125 of server 120. The user may then enter the search query via a graphical user interface provided by server 120.

[0024] The search query may take different forms, such as a fragment of text. The text fragment may be associated with a partial sentence, such as "Jane, I have to go because." Alternatively, the text fragment may be associated with a partial paragraph, such as "Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived, and so dedicated, can long endure. We are met on a great battle field of that war." While the description to follow will be described mainly in terms of completing sentences, the description is equally applicable to completing paragraphs.

[0025] Server 120 may perform a search for documents that contain the search query and retrieve the search results (act 320). For example, server 120 may search a corpus or repository

of documents to identify documents that include the text fragment of the search query as a phrase. In another implementation, server 120 may search for documents that also include synonyms of the word(s) in the search query. In either case, the documents may include documents stored by one or more servers, such as servers 120-140. Server 120 may optionally cap the number of documents included in the search results (e.g., server 120 may retrieve the top 100 documents). For each of these documents, server 120 may retrieve its title and text.

[0026] Server 120 may then determine whether there are sufficient search results (act 330). For example, server 120 may compare the number of search results retrieved with a threshold (e.g., five). When the number of search results is less than the threshold, the search results may not be adequate to satisfy the search query provided by the user. In this case, server 120 may form a shortened search query (act 340). For example, server 120 may drop one or more words from the search query.

[0027] Several techniques exist for determining what word(s) to drop. For example according to one implementation, server 120 may simply drop one or more words from the beginning or end of the search query. According to another implementation, server 120 may drop one or more words based on one or more symbols, such as a comma, semicolon, bracket, backslash, etc., contained in the search query. For example, if the search query includes a comma, then server 120 may drop everything before or after the comma. Server 120 may perform similar functions based on other symbols. According to yet another implementation, server 120 may analyze the structure of the search query to more intelligently drop one or more words. For example, server 120 may use a parse tree to identify parts of the search query. Server 120 may then drop one or more of these parts. In the sentence example provided above,

server 120 may shorten the search query to "I have to go because," dropping "Jane," from the search query.

[0028] Server 120 may then perform a search for documents that contain the shortened search query and retrieve the search results (act 320). As described above, server 120 may search a corpus or repository of documents to identify documents that include the shortened search query as a phrase. Server 120 may then again determine whether there are sufficient search results (act 330).

[0029] When there are sufficient search results (e.g., the number of search results is greater than or equal to the threshold), server 120 may scan the text of the documents in the search results to identify sentences that contain the search query (act 350). Server 120 may optionally locate periods within the documents to identify candidate sentences and then identify which of the candidate sentences include the search query. The search query may be included at the beginning or elsewhere within the identified sentences. Server 120 may give preference to a sentence that includes the search query at the beginning of the sentence over sentences where the search query occurs elsewhere. Server 120 may optionally discard sentences where the search query occurs more than once within the same sentences.

[0030] For each occurrence of the search query, server 120 may search left and right to determine the rough boundaries of the sentence containing the search query. For example, server 120 may look for periods (or other forms of punctuation) that typically precede and end a sentence. Server 120 may be programmed to ignore other typical occurrences of periods (and other forms of punctuation), such as when periods are used for initials, abbreviations, etc. Server 120 may optionally discard sentences that are missing punctuation and sentences that do not make sense (e.g., do not contain proper sentence structure).

[0031] Server 120 may then determine the sentence endings (also called "completions") associated with the identified sentences (act 360) (Fig. 3B). For example, server 120 may identify the word(s) that follow the text fragment of the search query until the end of the sentence. Server 120 may define a quality sentence ending as one that "ends properly," where "ends properly" is defined as: (1) the word(s) at the end make a better end of a sentence than they do a beginning of a sentence (e.g., year and pen); and (2) the last word is not in a list of bad endings (which may be maintained by server 120) (e.g., vs, dr, and aug).

[0032] To help in determining whether a word makes a better end of a sentence than a beginning of a sentence, a set of inverse document frequency (IDF) tables may be generated. IDF refers to a measure of a word's importance. In this case, two IDF tables may be generated. One table (hereinafter referred to as "start IDF table") may include uni-grams and bi-grams that are common at the start of sentences. The other table (hereinafter referred to as "end IDF table") may include uni-grams and bi-grams that are common at the end of sentences. To determine what is "common," a corpus of documents may be analyzed to identify the text that occurs around a period. Whether a word makes a better end of a sentence may be determined by analyzing the start and end IDF tables.

[0033] Server 120 may optionally trim and/or merge the sentence endings (act 370). When determining whether to trim a sentence ending, server 120 may consider the text and symbols included in the sentence ending. For example, server 120 may compare text of the sentence ending to entries in the start and end IDF tables to determine whether to cut the text. Server 120 may also consider symbols, such as a comma, semicolon, bracket, backslash, etc., when identifying what text to cut. In one implementation, server 120 may treat the dash separately, considering the text until the dash as a substring and ignoring the text after the dash. Server 120

may also disregard entire sentence endings that contain a colon (to avoid noise from message postings). Single word sentence endings may be considered when the word is significant (e.g., it is a common ending in the end IDF table). Based on the foregoing, server 120 may further consider a sentence ending that: (1) ends properly; and (2) does not separate a preposition (or possessive) from its object.

[0034] When determining whether to merge sentence endings, server 120 may search for sentence endings that overlap (i.e., sentence endings that have one or more words in common). Sentence endings may be merged based on their common parts. When merging sentence endings, server 120 may permit some small differences between them. For example, the sentence endings "has four legs and has a tail and barks" and "has four legs and a tail" may be merged to "has four legs and a tail."

[0035] Server 120 may optionally score the sentence endings (act 380). For example, server 120 may score the sentence endings by popularity. In other words, sentence endings that occur more often in the documents retrieved by the search may be scored higher than sentence endings that do not occur as often. Server 120 may alternatively, or additionally, score the sentence endings based on where the text fragment of the search query occurs within the identified sentences. In other words, the sentence endings corresponding to sentences where the text fragment of the search query occurs at the beginning of the sentences may be scored higher than sentence endings corresponding to sentences where the text fragment occurs elsewhere within the sentences. Server 120 may also penalize sentence endings for being too long, decreasing their scores. Server 120 may separately consider all of the sentence endings that were used to create a merged sentence ending when determining the score of that sentence ending.

[0036] Server 120 may present the sentence endings to the user (act 390). If the sentence endings were scored in some manner, server 120 may organize the sentence endings into a ranked list that it may provide to the user. In one implementation, server 120 may present an initial group of sentence endings to the user. The user may then be permitted to cycle through subsequent groups in a conventional manner.

[0037] Fig. 4 is a diagram of an exemplary ranked list 400 according to an implementation consistent with the principles of the invention. The exemplary ranked list 400 may include ranked items that each include a score 410 and a sentence ending (or "completion") 420. In this example, the user has provide a partial sentence of "I need to go now because." Server 120 provided various sentence endings that complete the partial sentence. In this example, the top-ranked sentence ending is "I have to get up early tomorrow."

[0038] In another implementation consistent with the principles of the invention, server 120 may provide sentence endings via a different interface. For example, server 120 may operate in conjunction with an application, such as a word processing application, an instant messenger application, an e-mail application, or another type of application via which documents (including messages) are prepared or edited. In any case, a server assistant, which may be in the form of executable code, such as a plug-in, an applet, a dynamic link library (DLL), or a similar type of executable object or process, resident on client 110, may operate to obtain the sentence endings from server 120. For example, the server assistant may notice text fragments that may require completion and communicate with server 120 to obtain the sentence endings. The server assistant may "notice" the text fragments by detecting them automatically to obtain the sentence endings on-the-fly or by detecting them when instructed by the user.

[0039] According to one implementation, the server assistant may automatically insert one of the sentence endings at the location of the user's cursor. For example, if the user types "I need to go because" and presses a special key, the server assistant may complete the sentence by automatically inserting one of the sentence endings. The user may then be permitted to view other possible sentence endings by pressing the special key again. Alternatively, subsequent sentence endings may be automatically presented after expiration of a possibly user-configurable amount of time. According to another implementation, the server assistant may present the sentence endings via a pop-up window, another type of interface, or a combination of interfaces (e.g., a first possible sentence ending may be automatically inserted, but subsequent sentence endings may be presented via a pop-up window).

CONCLUSION

[0040] Systems and methods consistent with the principles of the invention may automatically complete a fragment of text, such as a sentence or paragraph. The systems and methods may identify possible endings from text in web documents.

[0041] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, while series of acts have been described with regard to Figs. 3A and 3B, the order of the acts may be modified in other implementations consistent with the principles of the invention. Also, non-dependent acts may be performed in parallel. Further, while the acts of trimming and merging have been described as preceding the act of scoring, the scoring act may be performed prior to the trimming and/or merging acts.

[0042] Also, automatic paragraph completion has been described briefly. In one implementation, server 120 may provide a separate interface for paragraph completion. In another implementation, server 120 may provide the same interface for sentence and paragraph completion. When searching for paragraph endings, server 120 may also look for synonyms of the words provided in the search query. Server 120 may provide paragraph endings separately from or along with sentence endings. For example, server 120 may score the paragraph endings and the sentence endings and rank them based on their scores. It may be possible for server 120 to provide paragraph endings instead of sentence endings when server 120 finds no (or very few) good sentence endings for the search query.

[0043] Further, it has generally been described that server 120 performs most, if not all, of the acts described with regard to the processing of Figs. 3A and 3B. In another implementation consistent with the principles of the invention, one or more, or all, of the acts may be performed by client 110. For example, client 110 may obtain a text fragment and search documents local to client 110 (e.g., documents stored by client 110 and/or documents stored by a database accessible by client 110) to identify one or more documents that contain the text fragment. From these documents, client 110 may then identify potential sentence completions for the text fragment.